采用连续词袋模型(CBOW)的领域术语自动抽取研究*

姜　霖 [1,2]　王东波 [3]
[1](南京大学信息管理学院　南京 210023)
[2](江苏省数据工程与知识服务重点实验室　南京 210023)
[3](南京农业大学信息科学技术学院　南京 210095)

摘要:【目的】更准确便捷地完成术语词汇的自动抽取。【方法】利用 CBOW 模型计算构成术语的各个词部件的向量空间模型。通过词向量之间的余弦相似度衡量术语词汇内部各个词部件的关联度。利用 PageRank 算法计算候选词汇的领域代表性并排序, 通过阈值的设定, 抽取出更为具有领域代表性的术语词汇。【结果】在以自然语言处理领域内的论文摘要作为数据集的实验中取得较高的准确率和召回率。【局限】测试的数据训练集偏小, 而数据集的训练效果直接影响实验的效果。【结论】实验结果表明利用 CBOW 模型完成术语的抽取工作是一个较为合理、可行的方法。

关键词: 术语抽取　神经网络　CBOW 模型
分类号: TP18　G35

1　引　言

术语被定义为“特定专业领域中一般概念的词语指称”。许多专业领域的术语, 会伴随着学科的发展而产生动态的更新。新技术、新信息、新知识的产生会推动潜在的新术语词汇的出现。新术语被不停地引入到各个不同的学科领域中, 而旧术语则有可能逐渐消亡亦或是被赋予新的含义。术语和术语学这种动态变化的本质更加推动了术语处理技术的不断发展。

在术语的自动抽取中使用了很多自然语言处理技术, 如: 统计分析、词性标注、语义分析等。但同时, 自然语言处理应用中也需要与术语相关的信息来协助处理专业文档。比如: 机器翻译、自然语言生成、词典编纂、句法分析和自动文摘等。基于此, 实现高效、快速的术语自动抽取对于自然语言处理技术的发展有重要的意义。为了提高当前术语抽取的准确度, 本文提出一种基于词部件扩展算法和神经网络算法相结合的术语抽取方法, 利用神经网络的词向量计算方法构造词扩展部件的向量空间模型, 利用余弦相似度判断各个词扩展部件间的内部关联强度, 实现对术语候选词集的初步过滤, 最后结合 PageRank 算法, 统计候选集中各个词汇的领域代表性, 借此完成对领域术语词汇的精确抽取。

2　相关研究工作

术语的构成一般分为单词型术语和多词型术语。在冯志伟主持建设的“数据处理术语数据库”GLOT-C中, 词组型术语就占了75.17%。吴云芳等[1]研究发现词组型术语的比例是74%, 而单词型术语仅为26%。张榕[2]则分析了一个包含8 150条术语的数据库, 并通过分词工具统计了这些术语的词长分布特征, 其中包含2、3、4个单词的术语最多, 一共占总数的71.723%,

通讯作者: 王东波, ORCID: 0000-0002-9894-9550, E-mail: db.wang@njau.edu.cn。

*本文系南京农业大学人文社会科学研究基金项目“人文社会科学组块级汉英平行语料库构建及知识挖掘研究”(项目编号:SK2013023)和国家自然科学基金项目“基于 CSSCI 的句法级汉英平行语料库构建及知识挖掘研究”(项目编号:71303120)的研究成果之一。

而长度大于 6 的术语仅为 0.572%。李芸[3]对 56 609 条网络技术术语进行统计分析，发现单词型术语的比例为 7.7938%，而包含 2 到 6 个单词不等的词组型术语占据的比例为 89.7101%。因此本文将主要的研究对象设定为多词型术语。

单语的术语抽取方法主要分为三类:

(1) 基于语言规则的方法，即通过专家编写的术语词典和规则模板完成对术语的抽取[4-6]。该方法虽然精度较高，但编写规则依赖于语言环境和领域主题，难以实现移植。

(2) 基于统计特征的方法，即基于术语内部词之间黏着度较高的假设，利用统计特征实现术语抽取。目前在术语抽取中被成功使用的统计特征包括卡方检验、对数似然检验、互信息[7]和 C-Value/NC-Value[8]等。但是仅仅依靠术语内部黏着度效果却并不理想，为了能够大大提高准确率，加斯特森和卡茨在 1995 年利用一个词性过滤器过滤候选短语，这个过滤器只允许可能的“短语”的模式通过[9]。此外基于统计特征的方法还存在一些缺点，例如互信息算法很难排除语料中超低频词和超高频词的干扰，用来判断字词间的关联强度也存在缺陷，并且难以扩展到多词术语中。

(3) 基于机器学习的方法，即将术语抽取任务转化为分类问题或标注问题，借助决策树(DT)、支持向量机(SVM)[10]、隐 Markov 模型(HMM)、条件随机场(CRF)模型[11]等，但此类方法一般需要借助大量的人工标注语料。例如，章成志[12]提出一种多层术语度的一体化术语抽取方法，并采用句子术语度的概念，将术语所在句子的所有词语均作为训练特征，使用条件随机场识别术语，但该方法依赖大量训练数据。Lee 等[13]提出一种不依赖词典、以规则作为特征，通过 SVM 分类抽取术语的方法，但该方法的召回率偏低。

上述研究充分地利用了语法规则、统计方法两者的优点，大大提高了术语抽取的准确率。但是传统的监督学习算法，如利用 CRF 进行术语抽取，需要大量的人工标注以提高抽取的准确性，而使用 SVM 等无监督的抽取算法，又会带来超低频词的平滑，算法扩展性差等问题。基于此，本文采用基于 CBOW 模型(Continuous Bag-of-Words Model)的神经网络算法有效解决这些问题。

3 术语抽取研究框架与关键技术描述

3.1 算法设计和实现

图 1 为术语抽取实验的算法框架。算法输入为待抽取的文本文献，算法主要分为 4 个子函数层次，分别为：文献信息预处理层、语言模型抽取层、语义模型抽取层和领域代表性挖掘层。

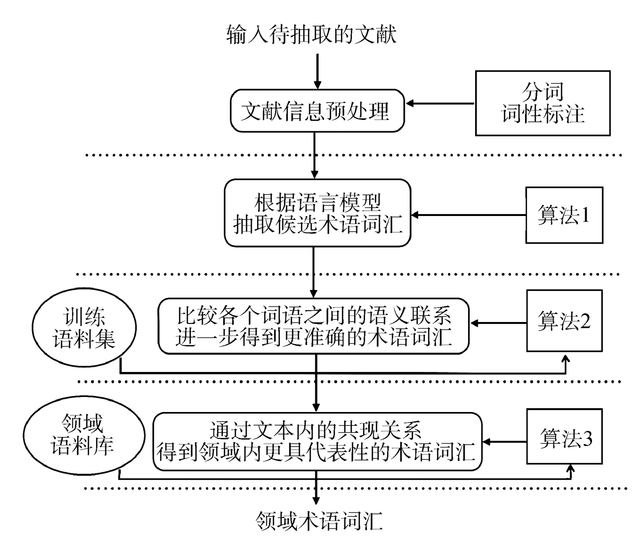


图 1 术语抽取算法框架

(1) 文献信息的预处理层

对于待抽取术语的文献资料，首先对其进行中文分词和词性标注处理，本文采用中国科学院计算技术研究所开发的 ICTCLAS 分词软件[14]对采集的汉语语料完成分词和词性标注工作。由于术语是一种能够表述具体概念的语言单元，隶属于实词的范畴。词性构成一般为名词、动词和形容词。针对术语词汇的这一特点，利用分词软件抽取出待抽取术语语料中的所有词性为名词、动词和形容词的词汇。

(2) 语言模型抽取层

算法 1 主要是利用预处理过的已经分过词、进行过词性标注的语料，利用语言模型提取出其中可能的候选术语语料集。为术语的抽取工作进行第一次初步的术语抽取和过滤。在实际操作过程中，人为添加了停用词表，例如“是”、“有”在文本语料中经常以动词的形式出现，但很少含有实际意义，也几乎不在术语词汇中出现，所以在抽取时利用词表将其去除，可以很好地提高抽取的效果。

(3) 语义模型抽取层

算法 2 主要是对初步筛选的候选术语进行进一步过滤，利用神经网络算法计算出每个词分量的语义向量，通过比较语义向量间的余弦相似度判断术语候选词中各个词扩展部件间的语义结合强度，以此得到更为准确的候选术语词汇结果。

若向量 $A=(A_1,A_2,\cdots,A_i,\cdots,A_n)$，$B=(B_1,B_2,\cdots,B_i,\cdots,B_n)$，则向量间余弦相似度的计算如下所示：

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}} \quad (1)$$

(4) 领域代表性挖掘层

算法 3 主要是为了评估得到的术语所具有的领域代表性，为此本文借鉴 PageRank 算法计算各个词部件在领域中的重要程度，并通过加和排序得到候选术语词集中最具有领域代表特征的术语词汇，从而使术语的抽取工作达到更好的效果。

3.2　关键技术描述

(1) 词部件扩展的使用原理

根据自然语言处理领域内术语的构成特点，将研究重点设置为多词型术语。假设某一特定领域，设 T 为该领域中一个多词型术语，构成它的词或者词缀设为 C，其构成的术语集合为 $SET_T=\{T|T=C_1\cdots C_{m-1},C,C_{m+1}\cdots C_n\}$，集合 SET_T 中的元素个数为 $n(n>1)$，构成术语词汇的每个词或者词缀称之为词扩展部件，由于大部分多词术语由两到三个词或者词缀组成，所以将研究的词部件的个数上限设置为 3，即 $4>n>1$。

(2) 术语语言模型运用原理

①术语是一种能够具体表述领域特征概念的语言单元，隶属于实词的范畴。所以词性构成一般为名词、动词和形容词。

②根据周浪[15]的统计分析，在两个词和三个词构成的术语中 Top5 的词法模式如表 1 所示，其中 N 代表名词，V 代表动词，A 代表形容词。

表 1　两词和三词术语 Top5 词法模式表

两词术语			三词术语		
序号	词性序列	示例	序号	词性序列	示例
1	N+N	“自然/n 语言/n”	1	N+N+N	“句法/n 标注/n 语料库/n”
2	V+N	“测度/v 空间/n”	2	N+V+N	“电路/n 交换/v 网络/n”
3	N+V	“机器/n 学习/v”	3	V+V+N	“并行/v 虚拟/v 机/n”
4	V+V	“编译/v 优化/v”	4	N+N+V	“自然/n 语言/n 处理/v”
5	A+N	“单调/a 函数/n”	5	V+N+N	“插/v 值/n 算法/n”

根据术语词汇中的词性构成特点，将本文重点研究的语言模型设定为表 1 中的 10 种。在实际术语词扩展部件的选取过程中，主要采用 ICTCLAS 分词软件对采集的汉语语料完成分词和词性标注工作。利用术语词汇的语言模型对候选术语完成初步提取工作。

(3) 术语词扩展部件的向量构建原理

本文中单词词向量的构建主要依据神经网络技术，采用 CBOW 模型[16-17]为词扩展部件构建词向量空间模型。

该模型的主要思想是在已知当前词 W_t 的上下文 $W_{t-2},W_{t-1},W_{t+1},W_{t+2}$ 的前提下预测 W_t。CBOW 模型的网络结构主要包括三层：输入层、投影层和输出层，如图 2 所示。

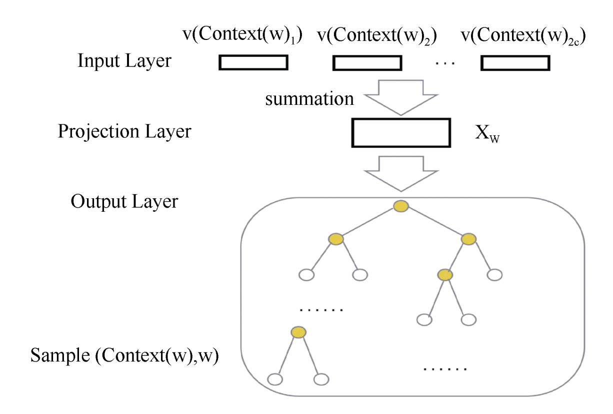


图 2　CBOW 模型网络结构

①输入层：包含 Context(w) 中 2c 个词的词向量 $v(Context(w)_1)$，$v(Context(w)_2)\ldots v(Context(w)_{2c})\in R^m$，m 表示词向量的长度。c 表示在词 w 的前后各取 c 个词。

②投影层：将 2c 个向量做求和累加，如下所示：

$$X_w = \sum_{i=1}^{2c} v(Context(w)_i) \in R^m \qquad (2)$$

③输出层：输出层对应一棵二叉树，以语料库中出现的词作为叶子节点，以各词在语料中出现的次数作为权值构造出霍夫曼树，树中叶子节点共 N(N=|D|)个，分别对应词典 D 中的词，非叶子节点 N–1 个(在图 2 中用深色标注的节点)。

实验主要采用基于 Hieratchical Softmax 的 CBOW 模型。目标函数通常为如下所示的对数似然函数：

$$\varsigma = \sum_{w \in c} \log p(w \mid Context(w)) \qquad (3)$$

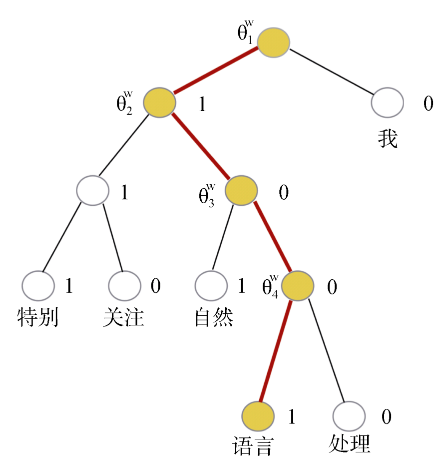


图 3　w="语言"时，相关记号示意

如图 3 中的霍夫曼树所示，对于词典中的任意词 w，霍夫曼树中必存有一条从根节点到词 w 对应节点的路径 p^w(且这条路径是唯一的)，路径 p^w 上存在 Y^w–1 个分支，将每个分支看成一个二分类，每次分类就产生一个概率，将这些概率乘起来就是所需的 p(w|Context(w))。使用随机梯度上升法，使目标函数最大化。这个神经网络中输出层的霍夫曼树的叶子节点上的向量为实验中使用的词向量。

(4) 基于语义的领域术语过滤原理

由于术语不同于一般的普通短语，是领域内具有表征概念的词语。所以术语词汇需要在所代表的领域内具有代表性。在领域文本集中，候选术语之间通过共同的上下文建立关联关系。在领域语料中的某个候选术语被其他候选术语关联越多，说明它越具有领域代表性，越有可能成为术语。为了能够体现出术语间的这种领域代表性，借鉴 PageRank 算法[18]，求出组成术语的各个词扩展部件在领域内的重要程度，以此进行排序，得到更具有领域代表性的术语词汇。

术语领域代表性的计算方法如下所示：

$$PR(A) = \frac{(1-d)}{N} + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \qquad (4)$$

其中，N 表示词总数，PR(A)表示词 A 的 PageRank 值，$PR(T_i)$表示和词 A 共现的 T_i 的 PageRank 值，$C(T_i)$表示词 T_i 和其他词共现的数量，d 为阻尼系数，取值范围为：0＜d＜1。在实际计算中每个词的初始 PageRank 值设为 1，阻尼系数设为 0.85。

4　实证研究

4.1　测试语料

实验利用网络爬虫从中国知网(CNKI)中以"自然语言处理"为检索主题，抽取 1 500 篇文献的中文摘要作为训练样本，并对其中的 500 篇完成人工标注作为对比实验样本，共得到 7 642 个术语词汇(主要为两词和三词词汇)，平均每篇摘要约 15 个术语词汇。人工标注提取的术语词汇如图 4 所示：

本文 分析 和 比较 了 几种 典型 的 线性 插值 方法 ，
着重 研究 了 它们 所 引发 的 词性 聚类 倾向 。
1、线性 插值 方法　　2、插值 方法
3、词性 聚类 倾向　　4、词性 聚类　　5、聚类 倾向

图 4　人工标注举例

4.2　评价指标

准确率和召回率是广泛用于信息检索和统计学分类领域的两个度量值，用来评价结果的质量。笔者在术语的抽取实验中也采用准确率、召回率和 F1 值作为评价参考，考量方法的实际使用效果。

$$准确率 = \frac{提取出的正确的信息条数}{提取出的信息条数} \qquad (5)$$

$$召回率 = \frac{提取出的正确的信息条数}{样本中的信息条数} \qquad (6)$$

$$F1值 = \frac{2 \times 准确率 \times 召回率}{准确率 + 召回率} \qquad (7)$$

4.3　结果与分析

以α、β作为判断词扩展部件间的关联强度和所具有的领域关键性的阈值设定。经过多次的反复对比实验，得到的结果具体如表 2 所示。从表 2 中可以直观地发现当α=0.6，β=0.015 时，F1 值相对最高，实验得到比较高的准确率和召回率。

表 2　实验结果

实验号	α	β	准确率	召回率	F1 值
1	0.5	0.0015	0.84	0.55	0.56
2	0.6	0.0015	0.87	0.60	0.71
3	0.7	0.0015	0.82	0.62	0.70
4	0.6	0.0014	0.83	0.53	0.65
5	0.6	0.0016	0.80	0.51	0.62

为了能够更加清楚地展示实验效果，示例将文献摘要"统计自然语言处理中，一个很复杂的问题是数据稀疏问题。主要有两种平滑方法解决：回退法和线性插值法。本文分析和比较了几种典型的线性插值方法，着重研究了它们所引发的词性聚类倾向。在此基础上，给出了 2 种改进的平滑方法。实验结果表明，改进的方法比原来的方法有更出色的平滑效果。"作为输入文本语料，从中抽取的具体术语结果如下：

自然语言处理	平滑方法	线性插值方法
语言处理	方法解决	词性聚类倾向
复杂问题	回退法	词性聚类
稀疏问题	线性插值	聚类性向
方法着重	插值法	平滑效果

可以发现，由于像"方法"这样的词汇在文本语料中的使用比较频繁而且和其他词汇都比较容易搭配，导致"方法解决"和"方法着重"在抽取过程中很难被过滤。在今后的实验方法上，应该更关注这样的词汇，希望通过其他的方法进一步提高术语抽取的准确度和召回率。另外由于设计的语言模型有限，所以例如"数据稀疏问题"这样的术语词汇未被抽取出来。随着语言模型的扩展添加和训练语料的扩展，实验的准确率和召回率可以得到进一步提升。

将实验方法与使用 N-Gram 模型建立向量空间模型作为 Baseline 对比，准确率有明显提高。使用神经网络模型较 N-Gram 模型主要有两个优势：

(1) 词语之间的相似性可以通过词向量体现。举例来说，如果 S1="计算机 软件"和 S2="电脑 软件"在语料中分别出现 1 000 次和 1 次，按照 N-Gram 模型，P(S1)一定大于 P(S2)，但是"计算机"和"电脑"是同义词并且承担相同的语法作用，所以 P(S1)应该与 P(S2)相似才更合理。在基于神经网络的算法中，P(S1)与 P(S2)是相近的，原因在于：在神经网络概率语言模型中假定"相似"的词对应的词向量也是相似的。并且概率函数关于词向量是光滑的，及词向量中的一个小变化对概率的影响也只是一个小变化。

(2) 基于词向量的 CBOW 模型自带平滑功能，由于 $p(w|\ Context(w))\in(0,1)$，不为零，所以不需要额外处理。此外与传统的基于 CRF 模型的抽取实验比较，在只进行少量的语料标注的情况下，基于 CBOW 模型的抽取实验在准确率和召回率上都明显优于 CRF 模型。

5　结　语

本文针对术语生成方式和结构特点，提出一种基于词部件扩展和神经网络相结合的术语抽取方法。与前人的研究相比，采用基于神经网络的 CBOW 模型构建基于语义的词部件向量空间模型，很好地解决传统互信息方法存在的词平滑问题，此外这种方法还可以避免大量的人工标注。通过利用向量间的余弦相似度衡量各个词扩展部件间的关联强度，通过关联强度的阈值设定完成术语词汇的抽取，可以有效地提高方法的可扩展性，加强对长术语的抽取效果。

综合而言，基于神经网络的词向量构建，便于以大量的领域语料库作为支撑，利用部件的领域聚合性特征完成术语抽取，并且可以得到较高的术语抽取召回率和准确率。但由于本次实验采集的数据量的限制，词扩展部件的向量计算和词领域代表性计算可能不精确，随着训练数据集的加大，实验得到的准确率和召回率可以进一步提升。

参考文献：

[1] 吴云芳，穗志方，邱利坤，等. 信息科学与技术领域术语部件描述[J]. 语言文字应用，2003(4): 34-39. (Wu Yunfang, Sui Zhifang, Qiu Likun, et al. The Approaches and Strategies to Describe the Term Component in Information Science and Technology [J]. Applied Linguistics, 2003(4): 34-39.)

[2] 张榕. 术语定义抽取、聚类与术语识别研究[D]. 北京：北京语言大学, 2006. (Zhang Rong. Research on Extraction and Clustering of Term Definition and Term Extraction [D]. Beijing: Beijing Language and Culture University, 2006.)

[3] 李芸. 信息科学和信息技术术语概念体系研究[D]. 北京：北京语言大学，2003. (Li Yun. Concept System of Terminology of Information Sciences and Information Technologies: A Preliminary Study [D]. Beijing: Beijing

Language and Culture University, 2003.)

[4] Bourigault D. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases [C]. In: Proceedings of the 14th Conference on Computational Linguistics. Association for Computational Linguistics, 1992: 977-981.

[5] Justeson J S, Katz S M. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text [J]. Natural Language Engineering, 1995, 1(1): 9-27.

[6] Ananiadou S. A Methodology for Automatic Term Recognition [C]. In: Proceedings of the 15th Conference on Computational Linguistics. Association for Computational Linguistics, 1994: 1034-1038.

[7] 张峰, 许云, 侯艳, 等. 基于互信息的中文术语抽取系统[J]. 计算机应用研究, 2005(5): 72-77. (Zhang Feng, Xu Yun, Hou Yan, et al. Chinese Term Extraction System Based on Mutual Information [J]. Application Research of Computers, 2005(5): 72-77.)

[8] Frantzi K, Ananiadou S, Mima H. Automatic Recognition of Multi-word Terms: The C-value/NC-value Method [J]. International Journal on Digital Libraries, 2000, 3(2): 115-130.

[9] Manning C D , Schutze H. 统计自然语言处理基础[M]. 范春法译. 第 4 版. 北京: 电子工业出版社, 2005: 95-97. (Manning C D, Schutze H. Foundations of Statistical Natural Language Processing [M]. Translated by Fan Chunfa. The 4th Edition. Beijing: Electronic Industry Press, 2005: 95-97.)

[10] Takeuchi K, Collier N. Use of Support Vector Machines in Extended Named Entity Recognition [C]. In: Proceedings of the 6th Conference on Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 1-7.

[11] Lafferty J D, McCallum A, Pereira F C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2001:282-289.

[12] 章成志. 基于多层术语度的一体化术语抽取研究[J]. 情报学报, 2011, 30(3): 275-285. (Zhang Chengzhi. Using Integration Strategy and Multi-level Termhood to Extract Terminology [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(3): 275-285.)

[13] Lee C M, Huang C K, Tang K M, et al. Iterative Machine-Learning Chinese Term Extraction [C]. In: Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012), Taipei, China. Berlin: Springer, 2012: 309-312.

[14] 刘克强. 2009 共享版 ICTCLAS 的分析与使用[J]. 科教文汇, 2009(22): 271, 280. (Liu Keqiang. The Analysing and Using of the 2009 Shared Version of ICTCLAS [J]. The Science Education Article Collects, 2009(22): 271, 280.)

[15] 周浪. 中文术语抽取若干问题研究[D]. 南京: 南京理工大学, 2009. (Zhou Lang. A Study on the Chinese Term Extraction [D]. Nanjing: Nanjing University of Science and Technology, 2009.)

[16] Mikolov T. Word2vec Code [CP/OL]. [2015-09-18]. http://word2vec.googlecode.com/svn/trunk/.

[17] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济, 2015, 25(2): 145-148. (Zhou Lian. Exploration of the Working Principle and Application of Word2vec [J]. Sci-Tech Information Development & Economy, 2015, 25(2): 145-148.)

[18] 罗刚. 解密搜索引擎技术实战[M]. 北京: 电子工业出版社, 2011: 73-74. (Luo Gang. Actual Explaining the Technologies of the Search Engine [M]. Beijing: Electronic Industry Press, 2011: 73-74.)

作者贡献声明:

姜霖: 提出研究思路, 进行实验与论文起草;
王东波: 改进研究方案, 采集数据;
姜霖, 王东波: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, 可通过电子邮件向作者索取, E-mail: 18205185622@163.com。

[1] 姜霖, 王东波. TermExtraction.rar. 术语自动抽取实验的程序代码.

[2] 姜霖, 王东波. WordTraining.rar. 训练语料.

收稿日期: 2015-09-06
收修改稿日期: 2015-11-03

Automatic Extraction of Domain Terms Using Continuous Bag-of-Words Model

Jiang Lin[1,2]　Wang Dongbo[3]
[1] (School of Information Management, Nanjing University, Nanjing 210023, China)
[2](Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)
[3](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Objective**] This study tries to extract domain terms more accurately and conveniently. [**Methods**] First, proposed a method using the CBOW model to build word vectors for each component of the terms. Then, applied the cosine similarity to calculate the internal correlation degree among each term's individual components. To get more representative terms, we used the PageRank algorithm to rank the candidates. [**Results**] We obtained high recall and precision rates using the paper abstacts in the field of natural language processing as the training pool. [**Limitations**] The training pool was relatively small, which might influence the results. [**Conclusions**] This study shows that CBOW model is a more appropriate method to extract terminologies.

Keywords: Terminology extraction　Neural network　Continuous Bag-of-Words Model

NISO 发布新版期刊文章标签集(JATS)标准

2016年1月7日，美国国家信息标准组织(NISO)宣布正式发布JATS的更新版本：JATS 1.1(Journal Article Tag Suite 1.1), ANSI/NISO Z39.96-2015。这个新的官方版本是ANSI/NISO Z39.96-2012(也称为JATS 1.0)的修订版本, JATS 1.0第一次发表时间是在2012年7月。JATS的目的是定义一系列的XML元素和属性，使得期刊文章的描述能以一种通用的格式进行，从而使得期刊内容之间的交换成为可能。这一标签系列是想要保护期刊知识内容，使其能独立于最初交付的形式，并且允许归档以便能捕捉到现有资源的结构和语义成分。除此之外，JATS标准还包括三种这类系列的实施方案，被称作标签集，这种标签集旨为期刊文章内容提供保存、发布和标记作者信息的模型。

"JATS 1.1建立在JATS 1.0的基础之上, JATS 1.0是美国国家医学图书馆(NLM)DTD 3.0版本的继承者，广泛应用于工业领域。"美国国家医学图书馆NCBI技术信息专家和NISO JATS常设委员会联合主席Jeffrey Beck指出, "JATS用于标记全球出版商出版的数以千计的期刊，并且JATS还在发展之中。"

"直到2015年2月，用户关于JATS 1.0的所有评论在JATS 1.1版本之中都已得到NISO JATS常设委员会的解决。所有的改动也能够与JATS 1.0版本相兼容，这意味着任何一个文件，只要对于JATS 1.0版本是有效的，则对于JATS 1.1版本同样是有效的。" Mulberry技术公司总裁和NISO JATS常设委员会的联合主席B. Tommie Usdin认为, "JATS的采用者信任JATS 1.1中的改进功能是完全稳定的，并且将如预期一样良好运行。"

"JATS 1.0是被ANSI批准的，并且在2012年由NISO发表。之后，该标准的更新是由ANSI所允准的维护程序所负责管理，这意味着在全新的标准被采用之前，所有的评论是由NISO JATS常设委员会修改和批准的。" NISO项目的副主任Nettie Lagace评论道, "这一常设委员会评价所有评论的可行性和优先顺序，并且做出适当的回复，这些回复现在可以通过NISO JATS的网站查找到，所以任何用户可以查找得到关于这些改变的所有历史信息。"

NISO JATS 1.1 标准有两种可利用版本，分别是 XML 文档格式和 PDF 格式，均可以在 NISO 网站上获取：http://www.niso.org/workrooms/journalmarkup。支持文档和 DTD 格式的 XML 模式、RELAX NG 和 W3C XML 模式格式可在 http://jats.nlm.nih.gov 中查找。

(编译自: http://www.niso.org/news/pr/view?item_key=15a111620077dd6d418deb3618f2c23dccf861b6)

(本刊讯)